

TOWARDS A SYSTEMATIC QUANTIFICATION OF THE SPEECH SPEED

Wainschenker Rubén, Doorn Jorge, Castro Marcela
INTIA - Facultad de Ciencias Exactas
Universidad Nacional del Centro de la Provincia de Buenos Aires
Paraje Arroyo Seco - Campus Universitario, (7000) Tandil – Argentina
{rfw, jdoorn, mcastro}@exa.unicen.edu.ar

Abstract

Although it is widely used in every day activities and in several professions such as locution, shorthand and stenography among others, the magnitude speed of the speech does not have a precise way of measurement. Such notion is fundamental in the context of the study of the typical behavior of a main set of language allophones as the basic components for a robust speech synthesis. No precise information was found for Spanish and less for the variant spoken in Uruguay and center and south of Argentina usually called Riverplatean Spanish.

This article presents a quantitative characterization of the intuitive notions of fast, slow and normal speech speed. The article reports a research with a strong experimental foundation since its conclusions were obtained from 120 texts uttered by different speakers at different speeds in a context free of any sort of conditioning. More than a half of the texts were collected from public sources and from people who never knew they were recorded for this purpose.

Keywords: speech speed, allophones, voice synthesis, voice parameters.

1. Introduction

The speech is one of the main signs of the human intelligence; it is naturally used to give and receive information. The voice is the acoustic fulfillment of the human language. Currently, several disciplines have taken it as the object of their study. Every science or art approaches the speech from a different point of view according to its objective ^[1].

From the acoustic point of view, voice may be seen as the result of a source of sound altered by a selective filter representing the vocal tract; the properties of the filter vary continuously during the process of speech production. These variations depend on the profile of the vocal tract, which in turn depends on the position of the articulating organs such as the tongue and lips.

The ability to produce sound does not imply the ability of “speech”. From the point of view of the voice synthesis, the simple concatenation of elemental sounds of a language leads to a poorly intelligible and unnatural sound. The generation of natural and continue voice is a permanent challenge in system applications that require synthesized voice. The currently known approaches only produce a remarkable artificial sound ^[2,3].

A characteristic seldom taken into account is the speed of the speech. A general model capable of producing synthesized voice should consider speech speed as one of their issues. On the other side, to know the speed of the speech

is an unavoidable previous step to understand the behavior of the different allophones when the speech speed varies. The current techniques used to contract or expand locutions (time compression and expansion or Speech Skimmer); pay little attention to the impact produced in every allophone by the global speed of the speech ^[4]. When listening to compressed or expanded locutions usually degradation in the understanding of the message occurs ^[5]. Since the synthesis of human voice tries to mimic the nature, its primary requirement is to know it as much as possible. Known values of speed of speech in Spanish language available before this study only quantify the speed produced under normal conditions. The goal of this paper is to make a contribution in this sense.

The experimental part of this study was carried out in the context of the Spanish spoken in Central and South regions of Argentina (including Uruguay) usually called Riverplatean Spanish. Even if it has a few variations in relations to other versions of Spanish, most of the conclusions obtained are applicable to the language as a whole.

2. Speech speed

The notion speed of the speech plays an important role in the behavior of the allophones of any language. Within the framework of human voice processed by computers specially when having as a target the development of a

robust algorithm capable of producing synthesized voice of good quality, this importance increases^[2,3]. On the other hand this sort of data is also useful when dealing with illnesses provoking defects in the speech^[6].

Navarro Thomas^[7] defines normal speed as the one produced when 205 words per minute (wpm) are emitted while Loprete^[1] establishes this notion when this figure ranges from 120 to 150. The context of both issues are different since the former corresponds to Castilian Spanish spoken in the middle of the XX century while the latter refers to the Riverplatean Spanish at the end of the same century. Other sources^[8] seem to confirm at least the low limit of the normal speech speed. For example in the shorthand courses a student passes when he or she is able to handle a talk at 120 wpm.

In other languages the information available does not contain precise data for speed of the speech. For example in English language reports characterize the normal speech as such produced when voice is uttered from 130 to 200 wpm^[9].

The lack of precision persists when the quality and the unit of measurement are taken into account, since different texts contain different number of allophones per word. The common sense and the experiences described below make clear that the allophones per time unit is more invariant with the text nature, specially in short texts.

To evaluate the appropriateness of a measurement technique it is convenient to take into account the basic principles of measurement theory, widely used in other disciplines. Five categories of measurement are usually accepted^[10,11].

1. Qualitative
2. Ordinal
3. By intervals
4. Proportional
5. Cardinal

The quality of the information obtained increases from qualitative to cardinal measurements. The first improvement gained using an ordinal measurement consists in guaranteeing a partial order among measured objects. The second consists in the existence of a total order while the existence of a reference point or zero of the scale characterizes a proportional measurement. Finally, the cardinal measurement requires a precise, repetitive and reliable unit of measurement.

All measurement performed along the study reported in this article were done in allophones per second. The unit allophone does not comply with the properties required for a cardinal measurement, but it is closer to it than measurements based in word per minute. Qualitative measurements should not be underestimated since the extension of their use is very wide, however it is useful to increase their information content by means of a comparative proportional measurement.

3. Experimental process

This article reports the results of a research about speech speed measurement based in the number of allophones per second.

	Allophones (a)	Lasting (s)	Speed (a/s)
T1	260	22.00	11.8
T2	298	30.00	9.9
T3	395	30.00	13.2
T4	570	30.00	19.00
T5	443	30.00	14.8
T6	397	30.00	13.2
T7	577	30.00	19.2
T8	335	20.00	16.7
T9	220	20.00	11.0
T10	218	21.00	10.4
T11	305	16.00	19.1
T12	95	28.00	3.4
T13	216	29.00	7.5
T14	413	28.00	14.8
T15	626	30.00	20.9
T16	903	31.70	28.5
T17	130	36.62	3.6
T18	427	29.70	14.4
T19	791	28.80	27.5
T20	61	31.00	2.0
T21	166	30.50	5.4
T22	364	30.00	12.1
T23	574	34.33	16.7
T24	826	35.34	23.4
T25	130	33.35	3.9
T26	381	31.08	12.3
T27	714	28.76	24.8
T28	445	31.36	14.2
T29	701	32.60	21.5
T30	488	33.00	14.8
T31	567	33.00	17.2
T32	356	31.53	11.3
T33	67	30.00	2.2
T34	543	40.00	13.6
T35	235	20.00	11.8
T36	310	25.00	12.4
T37	486	30.00	16.2
T38	216	30.00	7.2
T39	227	30.00	7.6
T40	708	30.00	23.6

Table 1. Number of allophones, overall duration and allophones per second of selected text

During the experiment, 30 different speakers uttered 120 texts at different speeds.

Half of them read a text knowing the nature of the experience and the other half were recorded from public domain sources such as radio and TV programs.

From a set of 120 available texts, 40 were chosen emphasizing the objective of having the most complete and uniform sample possible reducing any possible bias given by speakers, source, speed and others. The number of allophones, and the overall duration of the locution were measured for every selected text as showed in Table 1.

Text	Speed	Very Slow	Slow	Normal	Fast	Very Fast
20	2.0	31				
33	2.2	31				
12	3.4	31				
17	3.6	31				
25	3.9	31				
21	5.4	9	22			
38	7.2	4	24	3		
13	7.5	4	24	3		
39	7.6		22	9		
2	9.9		21	10		
10	10.4		7	24		
9	11.0		4	27		
32	11.3		5	26		
35	11.8		5	26		
1	11.8		5	26		
22	12.1		4	27		
26	12.3		5	26		
36	12.4			31		
3	13.2			28	3	
6	13.2			28	3	
34	13.6			28	3	
28	14.2			25	6	
18	14.4			25	6	
14	14.8			25	6	
5	14.8			20	11	
30	14.8			23	8	
37	16.2			8	23	
23	16.7			8	23	
8	16.8			5	26	
31	17.2			3	23	5
4	19.0				26	5
11	19.1			7	24	
7	19.2				21	10
15	20.9				21	10
29	21.5				20	11
24	23.4					31
40	23.6					31
27	24.8					31
19	27.5					31
16	28.5					31

Table 2. Comparison of speed speech quantitatively and qualitatively determined

The same group of 40 texts were measured qualitatively using an independent group of 31 listeners who, using their own criteria should assign one and only one of the following values: very slow, slow, normal, fast, very fast. None of the listeners had access to any other information about the text they did not even

know the label given to the text by other listeners.

The listeners were exposed to the recorded text at random, having an order for each listener. Based on this experimental study, it can be said that normal speech occurs when the speaker produces around 13 allophones per second, which corresponds to 150 wpm. A mean of 5.1 allophones per word was used for this comparison (see next section).

In long locutions a distorsive factor is introduced by the pauses between words, especially in slow locutions. In normal locutions these lapses are notoriously smaller and they vanish at high speed.

The experiment also gave quantitative data for very fast, fast, slow and very slow speech speed that could not be compared with other sources. The values obtained are presented in Table 2.

4. Results

In every row the text number is given along with its quantitative speech speed measured in allophones per second and the number of listeners that classified the text in the same group. The texts are sorted by the quantitative value of the speech speed to emphasize how good the qualitative groups mach quantitative data.

Qualitative group	Average (a/s)	Standard Deviation (a/s)
Very Slow	3.3	1.3
Slow	8.5	2.1
Normal	12.9	2.0
Fast	17.7	2.2
Very Fast	24.4	3.0

Table 3. Average and standard deviation of the speech speed

Table 3 holds the average and standard deviation values for the five qualitative groups, expressed in allophones per second (a/s). It can be said now that when the speech speed is (3.3 ± 1.3) a/s, (8.5 ± 2.1) a/s (12.9 ± 2.0) a/s (17.7 ± 2.2) a/s (24.4 ± 3.0) a/s the utterance is very slow¹, slow, normal, fast and very fast².

As it is shown in Table 4, using average and standard deviation a confidence interval may be defined in such a way that it is more probable that an arbitrary locution belongs to a

^{1,2} It is understood that in both ends one deviation are no useful since no other group were defined..

qualitative group than to the neighbor group. Other useful value is the speech speed at which the probability to belong to a group or to belong to the following group is the same. In other words, at which speech speed theoretically the 50% of the listener will choose one group and the 50% will choose another group. These values can be used as limits between qualitative groups. The Table 4 shows these limits.

	Very Slow	Slow	Normal	Fast	Very Fast
Average	3.3	8.5	12.9	17.7	24.4
Deviation	1.3	2.1	2.0	2.2	3.0
Limits	5.3	10.7	15.2	20.5	

Table 4. Limits between qualitative groups

The figure 1 shows the curves of density of probability corresponding to the five categories used along this study. The vertical axis is presented in arbitrary units to easy their understanding.

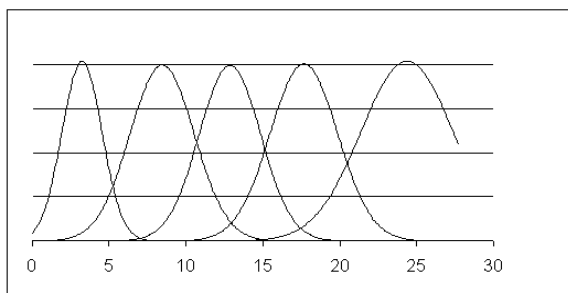


Figure 1. Theoretical distributions of probabilities for every qualitative group

Using the limits included in Table 4 it can be said that any text uttered at a speech speed below 5,3 a/s belongs to the group very slow, while if its speech speed is from 5.3 a/s to 10.7 a/s it can be considered as slow. Utterances from 10.7 a/s and 15.2 a/s correspond to normal locutions, while they can be considered fast if are emitted from 15.2 a/s to 20.5 a/s and finally they can be seen as very fast when more than 20.5 a/s are uttered. Roughly the limits between groups are 5, 10, 15, and 20 a/s.

The fact that the difference between these limits is almost the same value of 5 allophones per second gives an additional feeling of comfort with the data obtained besides the statistical proof.

All data presented in Table 1 through 4 were gathered including pauses between words. This was done in such way to make the comparisons with previous data easier. It can be argued that pauses distort the speed expressed in allophones

per second. The point of view supported in this article is to provide data about global speed of the speech. If pauses are removed, limits given in Table 4 become 5.3, 10.7, 15.2, 20.2.

5. Comparison with available data

As it was stated above, all the available quantitative data from bibliographic sources were expressed in words per minute. In order to compare words per minute with allophones per second, it was necessary to have a representative value for the magnitude allophones per word; so a new set of 30 texts was selected. Again different sources were chosen to decrease the influence of any possible bias. By counting words and allophones in those texts the figure of 5.1 allophones per word was obtained. The likelihood of this number is excellent in relation with the quality of the data available in bibliographic sources.

Converting the values of Table 4 to words per minute in the case of normal speech it can be said that it occurs when the locution varies between 120 and 180 words per minute. This value agrees with C. A. Loprete ^[1] who estimates the average speed for a locution between 120-150 wpm. Analogously, in his shorthand course, M. Vasallo and C. Fusca de Elías propose standards to take note of conversations beginning at 20 wpm increasing up to 102 wpm. These values will cover from very slow speech for beginners ending the course near to the upper limit of slow speech ^[8].

There are no available data about locutions at speeds different from normal, however the equidistance among the limits obtained leads to think that the obtained results are confidenceables in all groups.

6. Conclusions

The obtained results show that:

- The speed of the speech can be measured using as unit the number of allophones emitted per second.
- The daily practice is stable enough to characterize the speed of the speech at least in the five groups used in this paper.
- Reasonable relationship between the quantitative and the qualitative speech speed can be established.
- It is not known how the cognitive information about context influences the perception of the speed of the speech and how this factor affected the experiment. For

example is known for almost everybody that football relater speaks very fast, it is conceivable that this information might affected some of the collected data.

References

- [1] C. Loprete. 'El Lenguaje Oral: Fundamentos, Formas y Técnicas', Ed: Plus Ultra, 1984.
- [2] F. Casacuberta, E. Vidal. 'Reconocimiento Automático del Habla'. Ed: Marcombo, 1987, page 63-68.
- [3] R. Sproat. 'Multilingual Text-to-Speech Synthesis: The Bell Labs Approach', 1998, Lower Academic Publishers.
- [4] B. Arons. Speech Skimer: A System for Interactively Skimming Recorded Speech, ACM Transaction on Computer-Human Interaction, 1997, Vol. 4:1, page 3-38.
- [5] M. Covell, M. Withgott and M. Slaney. Mach1: Nonuniform Time-Scale Modification of Speech, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 1998, page 12-15.
- [6] C. Ferrer Riego, M. Hernandez-Díaz Huici. Obtención de un Índice Objetivo de Razón Lenta, Actas del VII Simposio de Comunicación Social, 2001, page 390-394.
- [7] T. Navarro. 'Manual de Pronunciación Española', Consejo Superior de Investigaciones Científicas, Madrid, 1950.
- [8] M. Vasallo, C. Fusca de Elias. 'Estenografía Vigente', Ed: Kapeluz, 1992, Vol. 2, page 166-169.
- [9] B. Arons. Techniques, Perception and Applications of Time-compressed Speech. Proceedings of 1992 Conference, American Voice I/O Society, 1992, page 169-177.
- [10] B. Klaassen Klaas. Electronic Measurement and Instrumentation, Cambridge University Press.,1996, page 1-15.
- [11] L. Briand, S. Morasca, V. Basili. Property-Based Software Engineering Measurement, IEEE Transactions on Software Engineering, 1996, Vol. 22:1, page 68-85.